Machine learning techniques

Machine learning types

1) Supervised learning
2) unsupervised learning
3) semi supervised learning
4) reinforcement learning

Supervised learning

A) continuous target variable → regression → House price prediction

B) categorical target variable → classification → Medical imaging

Unsupervised learning

No target variable available

A) Clustering → customer segmentation

B) association → market basket analysis

Semi supervised learning

A. Categorical target variable

→ classification → text classification

→ clustering → lane finding on GPS data

Reinforcement learning

A) categorical target variable

→ classification → optimized marketing

B) No target variable available

→ control → driver less cars

1) linear regression,ridge regression,lasso regression,

- continuous dependent variable
- linear relationship
- independent can be continuous or categorical
- data type - numerical data (if categorical convert to numerical)

2) logistic regression

- dependent variable should be binary categorical data
- independent can be continuous or categorical

3) decision tree, random forest,GBM, XGBoost, Light GBM, AdaBoost

- non linear supervised learning model
- can handle both classification and regression - Y
- independent variable can be numeric or categorical

4) cat boost

- optimized for data sets with categorical values
- does not require pre- processing like other models

5) svm

- good for both linear and non linear data
- performs well on categorical data
- works well on small data

6) k-nn

- suitable for small no of input features and large number of data points
- uses distant metrics to identity the closest neighbors
- used when labelled data is too expensive

7) GMM

- assumes data points are generated from mixture of Gaussian distributions

8) HIERARCHICAL clustering

- does not assume anything
- hierarchy of clusters by either merging or splitting clusters
- works well with both numerical and categorical

9) Fnn

- used for structured data where there is A clear mapping between input and output
- categorical features need to be encoded into numerical form
- can process on image data as well as text

10) CNN

- used for spatial data
- input - image / video

11) RNN

- data type - sequential
- used for time series, NLP, speech
- capture dependency across sequence elements

12) LSTM

- data type - sequential with long range dependencies

13)auto encoders

- used for unsupervised learning
- For dimensionality reduction and anamoly detection

14) seq2seq- paired input - output sequence

15) naive bayes - categorical and text data

For spam detection or sentiment analysis

Machine learning models

1) linear regression - data analysis technique that predicts the value of unknown data by using another related and known data

Example - real estate company wants to predict the selling prices of houses based on various factors. This can help then in setting prices for new houses and advising clients

2) ridge regression - statistical regularization technique. Also known as L2 . They add a penalty to the cost function that reduces overfitting.
It modifies the linear regression equation by adding a penalty term

3) lasso regression - statistical regularization technique. Also known as L1 regularization. Add absolute value of magnitude of the coefficient as a penalty term to the loss function where ridge uses squared magnitude of the coefficient as penalty

4) elastic net regression - linear regression technique that uses a penalty term as both l1 and l2 norm weighted by a parameter called alpha.
Useful for datasets with many predictors and multi collinearity

5) logistic regression - use to find answers to questions that have 2 or more finite outcomes. Appropriate when total count has an upper limit and initial growth is exponential.it is used for binary classification where we use sigmoid function that takes input as independent variables.

Example - marketing research firm use to predict the likelihood of a customer purchasing a product based on their age, income and education level

6) decision tree - non parametric supervised algorithm which is utilized for both classification and regression tasks
Represents a series of decisions and their possible consequence.

Example - decision tree is like a flow chart that helps a person to decide what to wear based on weathers condition

7) random Forest - combines the output of multiple decision trees to reach A single result and it can handle both classification and regression problems

Example - classifies whether an email is spam or not
For regression problems using features like size, no. Of bedroom, location, age to predict the selling price of houses based on these features

8) Gradient boosting machine - combines the prediction from multiple decision trees by building a model in a stage wise manner from weak learner's and improve by connecting the error of prev model

Example - can be used in financial forecasting to predict stock prices by combining multiple decision tree, each iteration refining the predictions based on the errors from earlier models

Both random forest and Gradient boosting are ensemble techniques

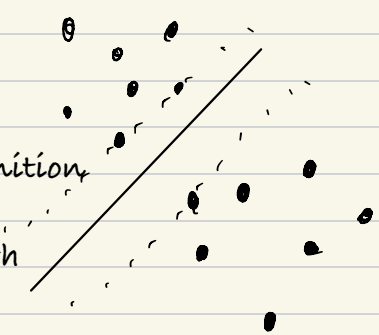
9) XGboost - is scalable distributed gradient boost where it provides parallel tree boosting and is one of the most used library.
Features - regularization, parallel processing, handle missing value, tree pruning, built in cv. Can be used in high stakes like fraud detection.

10) light GBM - faster training speed and higher efficiencies, it uses histogram based algorithm where it buckets continuous feature values into discrete bins which fasten the training procedure and also uses low memory usage

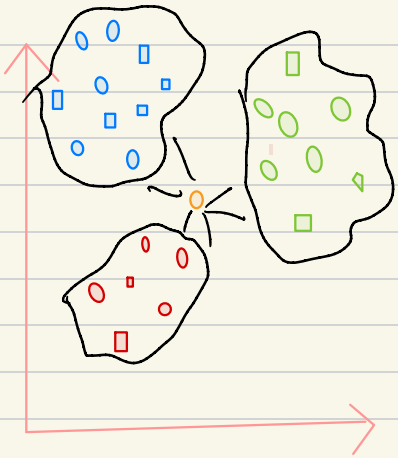
11) cat boost - designed for use on problems like regression having a very large number of independent features and it offers highest predictive accuracy but costs in computation
It works with categorical data and gradient boosting hence cat boost

12) Ada boost - combine multiple weak learner to create strong classifier. Each model is trained with a weighted dataset emphasising instances that previous model miss-classified

13) support vector machine - supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximises distance between each class in an n-d space
They distinguish between two classes by finding the optimal hyperlane that maximizes the margin between the closest data points of opposite Classes
They are trained using several functions
A) linear kernel function
B) quadratic
C) Gaussial radial basis
D) multilayer perception
Example - can be used for handwriting recognition, Intrusion detection, email classification Etc
It gives nigh quality results but difficult with big data and very slow for non linear sum



14) K- nearest neighbour - ML algorithm that uses proximity to compare one data point with a set of data it was trained on and has memorized to make predictions
Euclidean is the most commonly used one and the distance is sorted and most common among nearest neighbors is assigned
Small K value leads to low bias high variance.



15) principal component analysis - dimensionality reduction method that is often used to reduce the dimensionality of large data set by transforming a large set of variables into a smaller one that still contains most of the information

16) independent component analysis - techniques used to separate mixed signals into their independent sources

17) non negative matrix factorization - matrix V is factorized into W(basis matrix) and H(coefficient matrix). This method constraints to have non negative entries , helps in rank reduction

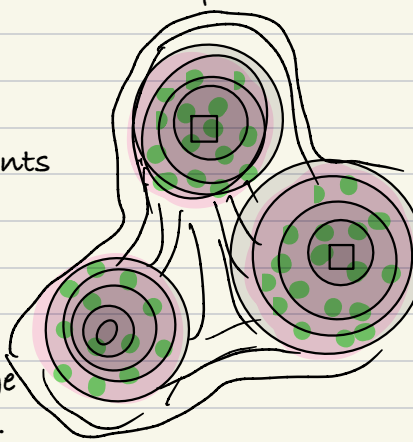
Example - image Processing where NMF decompose the image helping in feature extraction and noise reduction

18) Gaussian mixture model - soft clustering technique used in unsupervised learning to determine the probability that a given data point belongs to cluster. it estimates mean and covariance of the components

Application - anomaly detection, clustering and density estimation

19) hierarchical clustering - algorithm that groups similar objects into groups called clusters. The endpoint is set of clusters where each cluster is distinct from each other cluster and object within each cluster are broadly similar

Example - we have four cars that we can put into two clusters of car types: sedan and suv



20) mean shift clustering - shift each data point towards the highest density of distribution of points within a certain radius,

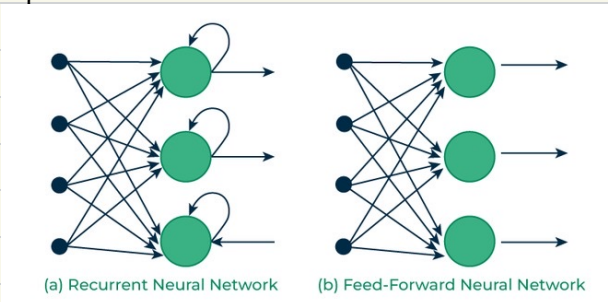
21) agglomerative clustering: begins with N groups each containing initially one entity and Then two most similar groups merge at each stage until there is a single group containing all data.

22) feed forward neural networks - information moves in only one direction ie forward from input nodes through the hidden nodes and to output nodes.
It consists of input layer, hidden layer and output layer

Example: used for image classification where image is an input and predicts the class label of the image

23) CNN - includes convolutional layers, max pooling layers and fully connected layers. Used to detect and classify object in an image

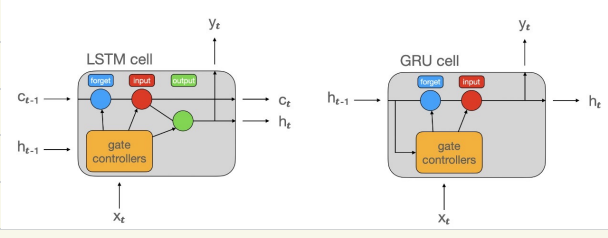
24) recurrent neural network - output from previous step is fed as input to the current step.



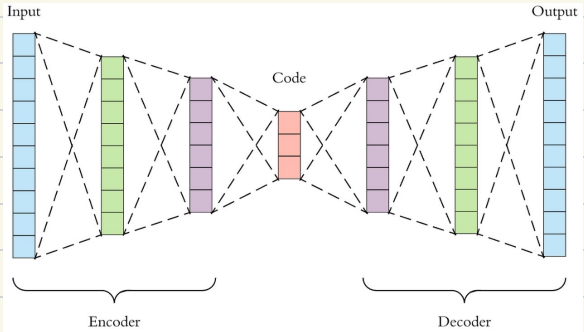
Example - can create a language translator with a RNN which analyses a sentence and correctly structure the words in different language

25) LSTM - stands for long short term memory with a strong ability to learn and predict sequential data.it has 3 gates which are input, forget and output gate. Can be used for speed recognition, time series prediction etc

26) gated recurrent unit - similar to LSTM where it is designed to model sequential data by allowing information to be selectively remembered or forgotten over time but with a simple architecture and fewer parameters



27) auto encoders - special type of neural network that is trained to cops its input to its output.
Image of handwritten digits is encoded into a lower dimensional representation later decoded back to an image
Can be used in medical imaging to Improve quality by removing noise And artifacts



28) variational auto encoder - enhanced form of an auto encoder that uses regularization techniques to overcome overfitting and ensure desirable properties

29) GAN - given a training set it learns to generate new data with the same statistics. Auto encoders just better the image where as GAN give completely new image

30) multi layer perceptron - consist of multiple layer of modes with each layer fully connected to next one. Used in NLP, image recognition and speech recognition

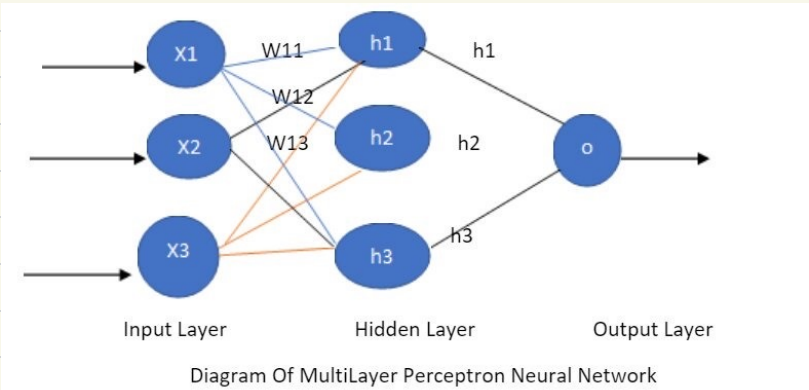
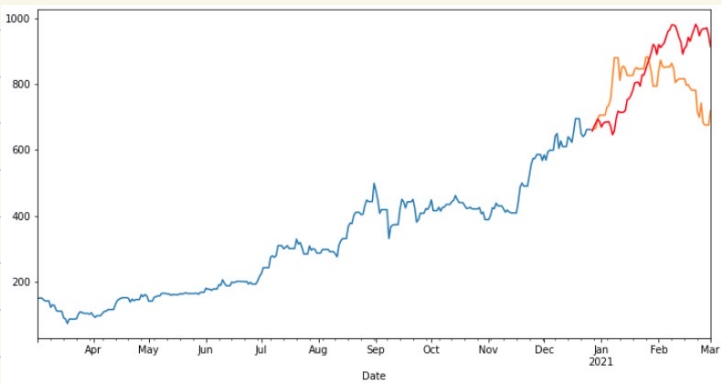


Diagram Of MultiLayer Perceptron Neural Network

31) seq2seq model - acts as an attention models that allows the decoder to focus on the most relevant parts of input sequence.it helps in boosting accuracy

32) word embedding model - way of representing words as vectors in multi dimensional space where the distance and direction between vectors reflect the similarity and relationship among the corresponding words.

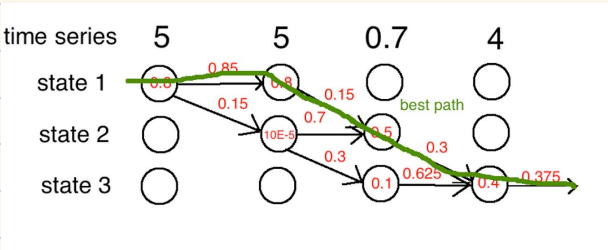
33) ARIMA - auto regressive integrated moving average
Analysis used for time series data
It predicts future values based on past values
Make use of lagged moving averages to smooth time series
Used in technical analysis to forecast future security prices



34) seasonalARIMA - extension of ARIMA that has seasonality in addition to non seasonal components

35) dynamic time warping - measure of similarity between two temporal sequences which may vary in speed.it replaces Euclidian distance by replacing one to one comparison with many to one composition

36) hidden Markov model for time series - can predict the future value of a time series based on its current and previous value. Used to capture underlying patterns in sequential data



37) naive bayes - probabilistic classifier based on applyingbayes theorem with naive assumptions between the feature
Assume the presence of a particular feature in class is independent of the presence of any other feature